Improving Large-Scale Image Retrieval using Geometric Weighting

Dmitry Sezganov and Moshe Porat

Abstract—The Bag-Of-Features (BOF) approaches are becoming central in large-scale image retrieval. The geometrical information is usually involved only in the post-processing spatial verification step usually implemented with the RANdom SAmple Consensus (RANSAC) algorithm. To enable visual search in real-time, RANSAC can be applied only to a relatively small number of top candidates due to its computational requirements. In this work, we propose an alternative method to perform accurate spatial verification with a significantly lower computational cost. Experimental results show that the proposed method outperforms the baseline BOF, and achieves similar performance as RANSAC based spatial verification, despite the major difference in complexity.

Keywords—Image Retrieval, Computer Vision and Pattern Recognition, Information systems & Applications, Scale-invariant Feature Transform (SIFT).

I. INTRODUCTION

VISUAL search in large-scale image and video databases has become an important task in recent years. It stems partly from the increasing use of mobile devices having integrated camera with decent image quality. Consequently, there is interest in performing visual search of specific objects in large-scale image databases, resulting in commercial products such as Google Goggles and others. There are also initial solutions for searching similar and near-duplicate images, e.g. Google Images and TinEye.

In this work we address the problem of large-scale visual search, in order to improve existing methods. We focus on a specific object recognition task in which the goal is to find all occurrences of a particular object in a very large image collection. Specifically, we are making accent on exploiting the geometric information in these methods. Our goal is to find all object instances despite possible changes in scale, location and viewpoint, as well as changes in illumination, background clutter and partial occlusion.

Inspired by the success of the bag-of-words approach for text retrieval, the majority of existing works represent images by Bag-Of-Features (BOF) models. The bags-of-features method is invariant to geometry, deformations and small viewpoint changes. It allows a very compact summary of image content and provides vector representation for sets of local features [12]. Although this approach has yielded good recognition results in practice, it still suffers from accuracy and scalability problems, which are very important in realistic computer vision applications.

A lot of effort has been recently put to improve retrieval accuracy such as solving quantization issue with soft quantization [10] and hamming embedding [5], expanding the search query to match additional images with query expansion [4], efficient large dictionary construction [9], [8] and inverted file compression [5]. However, most recent works completely ignore the geometry in an early stage of the retrieval. As a result, in large-scale image retrieval systems, the geometry is incorporated as a post-processing validation step, i.e., spatial verification [12], [7], [9], achieving the state-of-the-art results in terms of retrieval accuracy. Unfortunately, spatial verification is very computationally expensive and can be applied only to re-rank a small number of top image search results, requiring extensive memory operations to fetch local features of these images, which significantly slows down the search speed.

Other approaches try to capture visual word co-occurrence information, defining a new feature with increased discriminative power by grouping pairs of or multiple local features in a larger spatial neighborhood, e.g., the geometric min-Hash [3], doublets [11] i.e., pairs of visual words that co-occur within a local neighborhood. In [2], the authors propose to apply geometric hashing to local affine frames (LAFs) derived from boundaries of maximally stable extremal regions (MSER). Each local affine frame is described by a relative location of other local affine frames in its neighborhood represented in 6D hash table. The visual word co-occurrence methods are sensitive to the choice of representation of reference frame, while the number of combinations grows exponentially with the number of features.

Wang et al. [13] propose to apply a spatial clustering of salient points in the image, gathering the points of the same neighborhood and then characterizing each cluster with a BOF model. Spatial pyramid matching [6] encodes spatial information by enforcing the spatial distributions of local features belonging to the same category to be globally coherent. The method is not geometric invariant. Similarly in [1], local features of an image are projected to different directions or points to generate a series of ordered bag-of-features. Another way to group features is bundling features [14] that encode local spatial information in stable regions in inverted index. These approaches still need additional post-
processing to localize objects in the top retrieved images.

In [5], the authors add spatial constraints into the similarity measure between images by considering histograms of angle and scale differences of matched visual words. The new similarity measure provides a weak geometric consistency.

Visual phrase [15] encodes spatial information to inverted index considering local adjacency of visual words. However, the search precision is not as good as that of the RANSAC-based approach, since the encoded spatial information is too weak.

Unlike these studies, in this work we propose an alternative method to perform accurate spatial verification with a significantly lower computational cost. Our experimental results show that the proposed method outperforms the baseline BOF, following RANSAC based spatial verification.

This paper is organized as follows. In Section II we describe the new method for spatial verification. In Section III we describe and discuss our image retrieval framework. Then, comparable recognition performance results are shown in Section IV. Our conclusions and future work are presented in Section V.

II. VISUAL SEARCH WITH GEOMETRIC WEIGHTING

First, we define a zero ellipse with minor and major axeses $a = 1$, $b = 2$, located at the origin. Each elliptic feature is approximated by similarity transformation between a zero ellipse and this feature. The scale is estimated as a root square of the ellipse area. The transformation is defined as

$$A = \begin{bmatrix} sc & ss & tx \\ -ss & sc & ty \\ 0 & 0 & 1 \end{bmatrix},$$

where

$$sc = scale \cdot \cos(angle)$$
$$ss = scale \cdot \sin(angle).$$

Given a visual word, defined by a transformation $A$ (see Fig. 1) and the same visual word from another image defined by a transformation $B$, the transformation between the query image coordinate system and the other image coordinate system is

$$T = BA^{-1}.$$

All matching pairs of visual words having the same or nearly the same transformation $T$ are contributing their vote for a specific pose of the searched object. We need an efficient way to group nearly identical transformations.

For efficient grouping and comparison of multiple transformations we propose to quantize them. The quantization can be achieved through clustering with the Hough transform. This approach is discussed in the following section.

A. Transformation space voting

The pose of the object with respect to the query object is defined by a transformation from the query coordinate system to the potential image coordinate system. A set of possible solutions forms the transformation space. The dimensionality of this space equals the number of degrees of freedom in chosen family of transformations. Each point in the transformation space defines a pose of an object in respect to the query object.

We focus on the case where the transformation is similarity, allowing only rigid objects. The space of possible transformations is four-dimensional, having two dimensions for translation, one for scale and one for rotation. This is an approximation to the full 6 degrees-of-freedom affine model and also does not account for non-rigid deformations. This approximation together with noisy feature parameters imply large error bounds on the transformation space. Therefore, we use a large bin size for quantization to ensure high probability of collision for correct matching pairs. To overcome the problem of boundary effect in the bin assignment, we also add votes to all neighboring bins.

III. RETRIEVAL FRAMEWORK

A. Feature extraction and quantization

We use the same local features as in [9]. For each image we extract 128-D Scale-invariant Feature Transform (SIFT) descriptors, calculated for each detected affine-invariant Hessian region. Typically, in 1M pixel image there are about 3000 descriptors.

We use approximate k-means (AKM) clustering algorithm to build the visual dictionary with 1M visual words. All descriptors are quantized using dictionary centroids with ANN and then used in inverted file index. In the inverted index we store one entry per descriptor and not per word as in baseline BOF. However, in practice, for a very large visual dictionary, the memory requirements are equivalent to storing entries per visual word.

B. Searching with geometric weighting

Each indexed image is represented by a histogram of visual words. The histograms are typically very sparse and therefore can be efficiently indexed using an inverted file index. Given a
query image, the descriptors are extracted and quantized using the same dictionary. We can find the list of images in which all the given query words occur. This list is usually too big for a RANSAC-based spatial verification step, especially for very large image databases.

In order to avoid the post-processing spatial verification step we propose to integrate spatial information into BOF histogram representation. It is done by weighting each histogram bin according to the number of matching word pairs, sharing the same transformation. We calculate the similarity between a query image and each candidate image as follows:

1) For each query word $i$, extract all indexed entries containing the image ID, where this feature has been extracted from, its location, orientation and scale. The extracted words belonging to the same image form the histogram $H$.

2) For each extracted word, calculate the transformation that transforms a matching query word $i$ to that word.

3) Quantize all transformations using a 4D sparse array $B_k$, created for each candidate image $k$ as described previously. The corresponding quantization bin is incremented by 1, and also all the 80 neighboring bins are incremented by 0.5.

4) For each candidate image $k$, initialize with ones the weight vector $\mathbf{weight}_k$ of the same size like image histogram. For each matching word $j$ between the query image and image $k$ update the weight vector

$$\mathbf{weight}_k = B_k(T_j),$$

where $B_k(T_j)$ denotes the value of the bin where the transformation between matching words pair $j$ is quantized to.

5) For a query image and each candidate image update the histogram vector

$$\tilde{H}_k = H_k \cdot \mathbf{weight}_k / \|H_k \cdot \mathbf{weight}_k\|_1.$$  

6) Calculate the final score for each candidate image as

$$\text{score}_k = \|\tilde{H}_k - \tilde{H}_\text{query}\|_1.$$  

There are different ways to define similarity between two images. We use $l_1$ normalized histograms to calculate the similarity with $l_1$ distance measure.

### IV. EXPERIMENTAL RESULTS

#### A. Evaluation and dataset

The performance in all the retrieval experiments is measured using an average precision (AP), the area under precision-recall curve. The precision is defined as

$$\text{Precision} = \frac{TP}{TP + FP}$$

#### B. Baseline

As a baseline algorithm we choose the standard BOF approach with image histograms normalized $l_2$ and using $l_2$ distance with term frequency–inverse document frequency (TF-IDF) weighting. Other methods that compress inverted files, improve quantization to visual words, use additional post-processing steps are complimentary to our method and are not discussed here.

#### C. Comparison

Table I shows the retrieval accuracy on Oxford5K dataset, comparing our method to the baseline, defined above, and also the baseline followed by the RANSAC spatial verification step. Our approach performs significantly better than the baseline algorithm due to its built-in geometric verification nature. The result are similar to the baseline+RANSAC (we cite this results from [9]) but without the need for a computationally extensive RANSAC step. Our method requires only one pass over extracted words, compared to RANSAC that requires calculations for each word $N$ times, where $N$ is a number of iterations. Our spatial verification algorithm operates more than an order of magnitude faster than RANSAC spatial verification, allowing to apply it to more retrieved images. This can be especially important for very large databases when the precision of the top images is low. Some of the visual results are shown in Fig. 2. It can be seen from the results that the probability of collision for a group of correct matches is relatively low. It emphasizes the need for voting in neighboring bins for better recognition results.

Recall precision graphs are shown in Fig. 3. We found that the performance is better when not using the TF-IDF weighting, so it is omitted.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean average precision*</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>0.613</td>
</tr>
<tr>
<td>baseline + RANSAC</td>
<td>0.651</td>
</tr>
<tr>
<td>transformation voting</td>
<td>0.648</td>
</tr>
</tbody>
</table>

*See text for the numerical measure for mAP.

The precision is defined as

$$\text{Precision} = \frac{TP}{TP + FP}$$

and the recall

$$\text{Recall} = \frac{TP}{TP + TN}.$$  

Ideally, the precision should be 1 for all the recall values, resulting in average precision of 1 (the area under the curve). We calculate an average precision score for each test query, averaging them to get a mean average precision (mAP).

We use a publicly available image retrieval dataset called Oxford5k [9] that contains 5062 images with about 16M local features. It provides 55 test queries along with labeled desired results.

**Table I: Comparison of mAP for different methods on Oxford5K**

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean average precision*</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>0.613</td>
</tr>
<tr>
<td>baseline + RANSAC</td>
<td>0.651</td>
</tr>
<tr>
<td>transformation voting</td>
<td>0.648</td>
</tr>
</tbody>
</table>

*See text for the numerical measure for mAP.
Fig. 2 The matching visual words between query (left) and candidate image (right). The red lines denote extracted matches from the inverted file. Other colors denote groups of words having similar transformation.

Fig. 3 Recall precision graphs comparing baseline and proposed algorithms. The numbers in parentheses are average precision results. (a) Results for Oxford Query 1. (b) Results for Oxford Query 2.
V. CONCLUSIONS AND FUTURE WORK

We have proposed a new approach to image retrieval using an efficient spatial verification step. Each matched pair of features is assigned a transformation that it satisfies using a quantization algorithm. The transformations vote for the pose of the searched object. The major benefit of this new algorithm is the possibility to work with partial information enabling an approximate search and allowing other optimizations during the search phase. The results show that our method outperforms the standard BOF algorithm and gives comparable result as spatial verification based on RANSAC, while decreasing dramatically the computational time.

More evaluation is needed to determine the scalability of our approach when searching very large image collections (e.g., more than million images). In the future, we plan to examine the performance of geometric weighting combined with soft quantization method. We anticipate that geometric weighting will allow to discriminate wrong matches effectively, while the soft quantization will improve the overall recall.

ACKNOWLEDGMENT

This research was supported in part the Ollendorff Minerva Center. Minerva is funded through the BMBF.

REFERENCES